v1版本 解释

📦 Step 1: Exploratory Data Analysis (EDA)

1.0 导入数据和库

- 导入 `numpy`，`pandas`，`matplotlib`，`seaborn`，用于数据处理和可视化。
- 用 `sklearn.impute` 中的 `SimpleImputer` 做缺失值填补。
- 加载数据集 `custom_covid19.csv`，并通过 `df.head()` 和 `df.info()` 初步了解数据格式、列数、数据类型与缺失情况。

1.1 构造目标变量并分类特征

- 通过 DATE_DIED 字段是否为 `9999-99-99` 判断是否死亡，构造出新列 `DIED`（0 = 未死亡，1 = 死亡）。
- 按列的唯一值数量与数据类型将特征分为：
  - `cat_cols`：类别变量（<=20个唯一值，不含目标列和日期）
  - `num_cols`：数值变量（int/float 且不在 `cat_cols` 中）

1.2 缺失值处理

- 将所有特殊缺失值（97、98、99）统一替换为 `np.nan`。
- 分类变量使用最频繁值（mode）填充；数值变量用中位数填充，效率高于 KNN。
- 这样做可以避免模型训练时因缺失报错。

1.3 异常值检测

- 仅检测 `AGE` 的异常值，使用 IQR 方法（1.5 倍上下四分位距）。
- 打印上下界限和异常值数量，使用箱型图可视化，但不剔除，因为树模型对离群值较鲁棒。

1.4 各变量分布可视化

- 分别对 `AGE`、`SEX`、`PATIENT_TYPE`、`ICU` 等变量画图：
  - 数值型用 `histplot`（直方图）
  - 类别型用 `countplot`（柱状图）

- 目的是理解变量偏态和类分布特征。

1.5 相关性热图

- 计算所有数值特征的皮尔森相关系数。
- 用热力图展示：AGE 与多数变量相关性较弱（<0.25），TEST_RESULT 几乎不相关，提示 O2 和 O3 难度高。

1.6 人工特征选择 + 卡方检验

- 根据医学常识挑选了每个任务的重要变量列表 `important_features_O1/O2/O3`。
- 并对所有分类变量做了与 DIED 的卡方检验，列出显著变量（p<0.05）。

1.7 分组均值比较

- 对 DIED=0 和 DIED=1 两组分别计算特征均值并可视化，发现 ICU、INTUBED、PNEUMONIA 等在死亡组中平均更高。

1.8 类别分布

- 统计并可视化 DIED 的类别分布：死亡者仅占约 7%，说明样本严重不平衡，因此分类模型需加 class_weight='balanced'。

1.9 保存清洗后数据

- 将数值型但应为整数的列（如 AGE）统一转为 Int64 类型。
- 最终将清洗后的数据保存为 `custom_covid19_cleaned.csv`。

🎯 Step 2: Task O1 – 死亡预测（分类）

2.1 数据准备

- 使用 O1 任务挑选的 10 个变量作为特征，目标是 `DIED`。
- 划分训练集和验证集，stratify=y 确保两类分布一致。

2.2 建模管道构建

构建三种模型的 pipeline：

- Logistic Regression：带有缩放（StandardScaler）和 balanced 类权重
- Random Forest：不需要缩放，直接建模
- MLP：加上缩放，hidden_layer = (20,)

2.2.1 超参数调优

- 对 Logistic Regression 调整正则强度 `C`
- 对 MLP 调整隐藏层大小和 `alpha`
- 用 5 折交叉验证+f1-score 找最优模型
- RF 直接使用默认参数（n=200）

2.3 验证集评估

- 评估每个模型在验证集上的表现，指标包括：
  - F1, Recall：针对正类（死亡者）能力
  - MCC：处理不平衡更稳健
  - AUC：衡量预测概率的区分能力

2.4 可视化结果

- 使用混淆矩阵直观展示预测效果。

2.5 独立测试集测试

- 加载 `proj-test-data.csv` 和 `proj-test-class.csv`
- 直接对 `best_model` 测试，计算分类指标和 MCC。

2.6 保存最佳模型

- 将表现最好的模型（MCC最高的）保存为 `best_death_classifier.pkl`。

🎯 Step 3: Task O2 – 年龄预测（回归）

3.1 数据准备

- 特征为 9 个疾病类变量（0/1）
- 划分训练集和验证集，预处理器用 ColumnTransformer：OneHotEncoder + StandardScaler

3.2 模型构建

- 四种模型：Random Forest、HistGBR、Ridge、Linear
- HistGBR 是核心模型，训练快，表现稳定

3.3 RF 调参

- 用 GridSearchCV 调整 n_estimators 与 max_depth，5 折交叉验证，最优模型为 best_rf

3.4 模型拟合

- 对所有模型进行训练，准备统一评估。

3.5 验证集评估

- 指标包括：MAE, RMSE, R^2，外加 Pearson 相关性
- HistGBR 表现最优：MAE≈11.75，R^2≈0.2177

3.6 保存最佳模型

- 保存 HistGBR 为 `histgbr_age_predictor.pkl`

3.7 可视化预测 vs 实际

- 散点图 + 理想参考线

3.8 独立测试集评估

- 测试集上结果几乎与验证集一致，说明模型泛化良好。

3.9 附加交叉验证（仅在答辩时展示）

- 5 折交叉验证结果 MAE ±0.08，RMSE ±0.10，说明模型稳定。

🎯 Step 4: Task O3 – 死亡者年龄预测（回归）

4.1 数据准备

- 筛选 DIED=1 的死亡样本
- 特征为重要病症类变量，共 11 个
- 目标为 AGE
- 划分训练集与验证集，填补缺失

4.2 建模管道

- 用 RandomForest 和 GradientBoosting，参数为默认

4.3 验证评估

- 整体效果差：$R^2 \approx 0.10$，说明变量对年龄几乎无预测能力

4.4 交叉验证

- 进一步确认模型不稳定，误差大

4.5 保存模型

- 虽然模型不推荐，但仍保存 GradientBoosting 作为流程展示

4.6 测试集评估

- 在独立测试集中 R^2 为负，说明模型几乎无效

4.7 小结

- 模型表现极差是因为：
  - 样本少
  - 特征信息量低
  - 年龄本身与其它变量弱相关

🧠 Step 5: 特征总结与项目反思

5.1 重要特征对比

- 表格列出各任务的重要变量

5.2 特征影响总结

- O1：重症指标是死亡主因
- O2：疾病与年龄弱相关
- O3：病症特征预测年龄失败

5.3 项目总结

- 完整走了一遍机器学习工作流
- O1 模型效果最好；O2 有一定误差；O3 无法建模
- 工具使用辅助但理解为主，收获大